Fixed Points of Generalized Approximate Message Passing with Arbitrary Matrices

Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson Fletcher, Volkan Cevher

***Abstract*—The estimation of a random vector $\mathbf{x}$ with independent components passed through a linear transform followed by a componentwise (possibly nonlinear) output map arises in a range of applications. Approximate message passing (AMP) methods, based on Gaussian approximations of loopy belief propagation, have recently attracted considerable attention for such problems. For large random transforms, these methods exhibit fast convergence and admit precise analytic characterizations with testable conditions for optimality, even for certain non-convex problem instances. However, the behavior of AMP under general transforms is not fully understood. In this paper, we consider the generalized AMP (GAMP) algorithm and relate the method to more common optimization techniques. This analysis enables a precise characterization of the GAMP algorithm fixed-points that applies to arbitrary transforms. In particular, we show that the fixed points of the so-called max-sum GAMP algorithm for MAP estimation are critical points of a constrained maximization of the posterior density. The fixed-points of the sum-product GAMP algorithm for estimation of the posterior marginals can be interpreted as critical points of a certain mean-field variational optimization.**

***Index Terms*—Belief propagation, ADMM, variational optimization, message passing.**

I. Introduction

Consider the constrained optimization problem

$$(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}) := \arg\min_{\mathbf{x},\mathbf{z}} F(\mathbf{x},\mathbf{z}) \quad \text{s.t. } \mathbf{z} = \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{z} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times n}$ and the objective function admits a decomposition of the form

$$\begin{aligned} F(\mathbf{x},\mathbf{z}) &:= f_x(\mathbf{x}) + f_z(\mathbf{z}) \\ f_x(\mathbf{x}) &= \sum_{j=1}^{n} f_{x_j}(x_j), \quad f_z(\mathbf{z}) = \sum_{i=1}^{m} f_{z_i}(z_i), \end{aligned} \tag{2}$$

for scalar functions $f_{x_j}(\cdot)$ and $f_{z_i}(\cdot)$. One example where this optimization arises is the estimation problem in Fig. 1. Here, a random vector $\mathbf{x}$ has independent components with densities $p_{x_j}(x_j)$, and passes through a linear transform to yield an output $\mathbf{z} = \mathbf{A}\mathbf{x}$. The problem is to estimate $\mathbf{x}$ and $\mathbf{z}$ from measurements $\mathbf{y}$ generated by a componentwise conditional density $p_{y_i|z_i}(y_i|z_i)$. Under this observation model, the vectors $\mathbf{x}$ and $\mathbf{z}$ will have a posterior joint density given by

$$p_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\mathbf{x},\mathbf{z}|\mathbf{y}) = [Z(\mathbf{y})]^{-1} e^{-F(\mathbf{x},\mathbf{z})} \mathbb{1}_{\{\mathbf{z}=\mathbf{A}\mathbf{x}\}}, \tag{3}$$

where $F(\mathbf{x},\mathbf{z})$ is given by (2) when the scalar functions are set to the negative log densities:

$$f_{x_j}(x_j) = -\log p_{x_j}(x_j), \quad f_{z_i}(z_i) = -\log p_{y_i|z_i}(y_i|z_i).$$

Note that in (3), $F(\mathbf{x},\mathbf{z})$ is implicitly a function of $\mathbf{y}$, $Z(\mathbf{y})$ is a normalization constant and the term $\mathbb{1}_{\{\mathbf{z}=\mathbf{A}\mathbf{x}\}}$ imposes the linear constraint that $\mathbf{z} = \mathbf{A}\mathbf{x}$. The optimization (1) in this case produces the *maximum a posteriori* (MAP) estimate of $\mathbf{x}$ and $\mathbf{z}$. In statistics, the system in Fig. 1 is sometimes referred to as a generalized linear model [1], [2] and is used in a range of applications including regression, inverse problems, and filtering. Bayesian forms of compressed sensing can also be considered in this framework by imposing a sparse prior for the components x_j [3]. In all these applications, one may instead be interested in estimating the posterior marginals $p(x_j|\mathbf{y})$ and $p(z_i|\mathbf{y})$. We relate this objective to an optimization of the form (1)–(2) in the sequel.

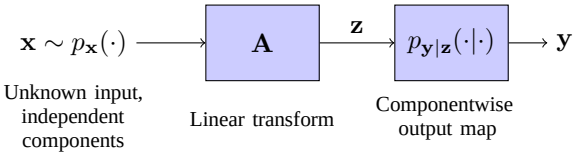

Fig. 1. System model: The GAMP method considered here can be used for approximate MAP and MMSE estimation of $\mathbf{x}$ from $\mathbf{y}$.

Most current numerical methods for solving the constrained optimization problem (1) attempt to exploit the separable structure of the objective function (2) either through generalizations of the iterative shrinkage and thresholding (ISTA) algorithms [4]–[10] or alternating direction method of multipliers (ADMM) approach [11]–[20]. There are now a large number of these methods and we provide a brief review in Section II.

However, in recent years, there has been considerable interest in so-called approximate message passing (AMP) methods based on Gaussian and quadratic approximations of loopy belief propagation in graphical models [21]–[26]. The main appealing feature of the AMP algorithms is that for certain large random matrices $\mathbf{A}$, the asymptotic behavior of the algorithm can be rigorously and exactly predicted with testable conditions for optimality, even for many non-convex instances. Moreover, in the case of these large, random matrices, simulations appear to show very fast convergence of AMP

S. Rangan (email: srangan@poly.edu) is with the Department of Electrical and Computer Engineering, Polytechnic Institute of New York University, Brooklyn, NY.

P. Schniter (email: schniter@ece.osu.edu) is with the Department of Electrical and Computer Engineering, The Ohio State University.

E. Riegler (email: erwin.riegler@tuwien.ac.at) is with the Institute of Telecommunications, Technische Universität, Wien.

A. K. Fletcher (email: afletcher@ucsc.edu) is with the Department of Electrical and Computer Engineering, University of California, Santa Cruz.

V. Cevher (email: volkan.cevher@epfl.ch) is with Ecole Polytechnic, Lausanne, Switzerland

methods when compared against state-of-the-art conventional optimization techniques.

However, despite recent extensions to larger classes of random matrices [27], [28], the behavior of AMP methods under general $\mathbf{A}$ is not fully understood. The broad purpose of this paper is to show that certain forms of AMP algorithms can be seen as variants of more conventional optimization methods. This analysis will enable a precise characterization of the fixed points of the AMP methods that applies to arbitrary $\mathbf{A}$.

Our study focuses on a generalized AMP (GAMP) method proposed in [26] and rigorously analyzed in [29]. We consider this algorithm since many other variants of AMP are special cases of this general procedure. The GAMP method has two common versions: max-sum GAMP for the MAP estimation of the vectors $\mathbf{x}$ and $\mathbf{z}$ for the problem in Fig. 1; and sum-product GAMP for approximate inference of the posterior marginals.

For both versions of GAMP, the algorithms produce estimates $\mathbf{x}$ and $\mathbf{z}$ along with certain quadratic terms. Our first main result (Theorem 1) shows that the fixed points $(\widehat{\mathbf{x}}, \widehat{\mathbf{z}})$ of the max-sum GAMP are critical points of the optimization (1). In addition, the quadratic terms can be considered as approximations of the diagonal Hessian of the objective function. For sum-product GAMP, we provide a variational interpretation of the algorithm's fixed points. Specifically, we show (Theorem 2) that the algorithm fixed points can be interpreted as means and variances of a certain Gaussian mean-field approximation of the posterior distribution. The results are compared against the well-known characterization of fixed points of standard loopy BP [30]–[33].

Algorithm 1 Generalized Approximate Message Passing (GAMP)

Require: Matrix $\mathbf{A}$, functions $f_x(\mathbf{x})$, $f_z(\mathbf{z})$. and algorithm choice MaxSum or SumProduct.

1: $t \leftarrow 0$
2: Initialize $\mathbf{x}^t$, $\boldsymbol{\tau}_x^t$
3: $\mathbf{s}^{t-1} \leftarrow 0$
4: $\mathbf{S} \leftarrow |\mathbf{A}|^2$ (componentwise magnitude squared)
5: **repeat**
6: {Output node update}
7: $\boldsymbol{\tau}_p^t \leftarrow \mathbf{S}\boldsymbol{\tau}_x^t$
8: $\mathbf{p}^t \leftarrow \mathbf{A}\mathbf{x}^t - \mathbf{s}^{t-1}\boldsymbol{\tau}_p^t$
9: **if** MaxSum **then**
10: $\mathbf{z}^t \leftarrow \arg\min_{\mathbf{z}} F_z(\mathbf{z}, \mathbf{p}^t, \boldsymbol{\tau}_p^t)$
11: $\boldsymbol{\tau}_z^t \leftarrow \boldsymbol{\tau}_p^t/(1 + \boldsymbol{\tau}_p^t \partial^2 f_z(\mathbf{z}^t)/\partial z^2)$
12: **else if** SumProduct **then**
13: $\mathbf{z}^t \leftarrow \mathbb{E}(\mathbf{z}|\mathbf{p}^t, \boldsymbol{\tau}_p^t)$
14: $\boldsymbol{\tau}_z^t \leftarrow \text{var}(\mathbf{z}|\mathbf{p}^t, \boldsymbol{\tau}_p^t)$
15: **end if**
16: $\mathbf{s}^t \leftarrow (\mathbf{z}^t - \mathbf{p}^t)/\boldsymbol{\tau}_p^t$
17: $\boldsymbol{\tau}_s^t \leftarrow 1/\boldsymbol{\tau}_p^t - \boldsymbol{\tau}_z^t/(\boldsymbol{\tau}_p^t)^2$
18:
19: {Input node update}
20: $\boldsymbol{\tau}_r^t \leftarrow 1/(\mathbf{S}^T\boldsymbol{\tau}_s^t)$
21: $\mathbf{r}^t \leftarrow \mathbf{x}^t + \boldsymbol{\tau}_r^t \mathbf{A}^T \mathbf{s}^t$
22: **if** MaxSum **then**
23: $\mathbf{x}^{t+1} \leftarrow \arg\min_{\mathbf{x}} F_x(\mathbf{x}, \mathbf{r}^t, \boldsymbol{\tau}_r^t)$
24: $\boldsymbol{\tau}_x^{t+1} \leftarrow \boldsymbol{\tau}_r^t/(1 + \boldsymbol{\tau}_r^t \partial^2 f_x(\mathbf{x}^t)/\partial x^2)$
25: **else if** SumProduct **then**
26: $\mathbf{x}^{t+1} \leftarrow \mathbb{E}(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
27: $\boldsymbol{\tau}_x^{t+1} \leftarrow \text{var}(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
28: **end if**
29: **until** Terminated

II. Review of GAMP and Related Optimization Methods

A. Generalized Approximate Message Passing

Graphical model methods [34] are a natural approach to the optimization problem (1) given the separable structure of the objective function (2). However, traditional graphical model techniques such as loopy belief propagation (loopy BP) generally require that the constraint matrix $\mathbf{A}$ is sparse. Approximate message passing (AMP) refer to a class of Gaussian and quadratic approximations of loopy BP that can be applied to dense $\mathbf{A}$. AMP approximations of loopy BP originated in CDMA multiuser detection problems [35]–[37] and have received considerable recent attention in the context of compressed sensing [21]–[26], [38]. The Gaussian approximations used in AMP are also closely related to expectation propagation techniques [39], [40].

In this work, we study the so-called generalized AMP (GAMP) algorithm [26] and rigorously analyzed in [29]. The procedure is shown in Algorithm 1 and produces a sequences of estimates $\mathbf{x}^t$ and $\mathbf{z}^t$ along with what we will call *quadratic terms* $\boldsymbol{\tau}_x^t, \boldsymbol{\tau}_z^t, \ldots$. We focus on two variants of the GAMP algorithm: *max-sum GAMP* and *sum-product GAMP*.

In the max-sum version of the algorithm, the outputs $(\mathbf{x}^t, \mathbf{z}^t)$ represent a sequence of estimates of the solution to the optimization problem (1), or equivalently the MAP estimates for the posterior (3). Each iteration of the algorithm involves computing minimizations shown in lines 10 and 23 whose objective functions are given by

$$F_x(\mathbf{x}, \mathbf{r}, \boldsymbol{\tau}_r) \;:=\; f_x(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{r}\|^2_{\boldsymbol{\tau}_r}, \tag{4a}$$

$$F_z(\mathbf{z}, \mathbf{p}, \boldsymbol{\tau}_p) \;:=\; f_z(\mathbf{z}) + \frac{1}{2}\|\mathbf{z} - \mathbf{p}\|^2_{\boldsymbol{\tau}_p}, \tag{4b}$$

where we use the notation that, for any vectors $\mathbf{v}$ and $\boldsymbol{\tau} \in \mathbb{R}^r$ where $\boldsymbol{\tau}$ has positive components:

$$\|\mathbf{v}\|^2_{\boldsymbol{\tau}} := \sum_{i=1}^{r} \frac{|v_i|^2}{\tau_i}.$$

Since the objective function has a separable form (2), it can be verified that the minimizations in lines 10 and 23 can be performed componentwise:

$$z_i^t = \text{prox}_{\tau^t_{p_i} f_{z_i}}(p_i^t), \quad x_j^{t+1} = \text{prox}_{\tau^t_{r_j} f_{x_j}}(r_j), \tag{5}$$

where $\text{prox}_f(\cdot)$ is the so-called proximal operator:

$$\text{prox}_f(v) := \arg\min_{u \in \mathbb{R}} f(u) + \frac{1}{2}|u - v|^2. \tag{6}$$

In this way, the max-sum GAMP algorithm reduces the vector-valued optimization (1) to a sequence of scalar optimizations.

Similarly, it can be verified that the quadratic terms in lines 11 and 24 depend on the derivative of the proximal operator:

$$\tau_{z_i}^t = \tau_{p_i}^t \operatorname{prox}'_{\tau_{p_i}^t f_{z_i}}(p_i^t), \quad \tau_{x_j}^{t+1} = \tau_{r_j}^t \operatorname{prox}'_{\tau_{r_j}^t f_{x_j}}(r_j). \quad (7)$$

For the sum-product GAMP algorithm, the outputs $(\mathbf{x}^t, \mathbf{z}^t)$ represent the posterior means for the density (3), or equivalently, the minimum mean-squared error (MMSE) estimates of $(\mathbf{x}, \mathbf{z})$ given $\mathbf{y}$. As discussed in [26], the algorithm also provides estimates of the posterior marginals. The expectations and variances in lines 13, 14, 26 and 27 of the algorihtm are to be taken with respect to the probability density functions:

$$p(\mathbf{x}|\mathbf{r}, \boldsymbol{\tau}_r) \propto \exp[-F_x(\mathbf{x}, \mathbf{r}, \boldsymbol{\tau}_r)] \quad (8a)$$
$$p(\mathbf{z}|\mathbf{p}, \boldsymbol{\tau}_p) \propto \exp[-F_z(\mathbf{z}, \mathbf{p}, \boldsymbol{\tau}_p)], \quad (8b)$$

where $F_x(\cdot)$ and $F_z(\cdot)$ are defined in (4). Under the separability assumption (2), these density functions factor as

$$p(\mathbf{x}|\mathbf{r}, \boldsymbol{\tau}_r) \propto \prod_{j=1}^n \exp\left[-f_{x_j}(x_j) - \frac{|x_j - r_j|^2}{2\tau_{r_j}}\right] \quad (9a)$$

$$p(\mathbf{z}|\mathbf{p}, \boldsymbol{\tau}_p) \propto \prod_{i=1}^m \exp\left[-f_{z_i}(z_i) - \frac{|z_i - p_i|^2}{2\tau_{p_i}}\right]. \quad (9b)$$

Thus, the expectations and variance computations can be computed componentwise and the sum-product GAMP algorithm reduces the vector-valued estimation problem to a sequence of scalar estimation problems.

B. Iterative Shrinkage and Thresholding Algorithm

The goal in the paper is to relate the GAMP method to more conventional optimization techniques. One of the more common of such approaches is a generalization of the Iterative Shrinkage and Thresholding Algorithm (ISTA) shown in Algorithm 2 [4]–[8].

Algorithm 2 Iterative Shrinkage and Thresholding Algorithm (ISTA)

Require: Matrix $\mathbf{A}$, scalar c, functions $f_x(\cdot)$, $f_z(\cdot)$.
1: $t \leftarrow 0$
2: Initialize $\mathbf{x}^t$.
3: **repeat**
4: $\quad \mathbf{z}^t \leftarrow \mathbf{A}\mathbf{x}^t$
5: $\quad \mathbf{q}^t \leftarrow \partial f_z(\mathbf{z}^t)/\partial \mathbf{z}$
6: $\quad \mathbf{x}^{t+1} \leftarrow \arg\min_{\mathbf{x}} f_x(\mathbf{x}) + (\mathbf{q}^t)^T\mathbf{A}\mathbf{x} + (c/2)\|\mathbf{x} - \mathbf{x}^t\|^2$
7: **until** Terminated

The algorithm is built on the idea that, at each iteration t, the second cost term in the minimization $\arg\min_{\mathbf{x}} f_x(\mathbf{x}) + f_z(\mathbf{A}\mathbf{x})$ specified by (1) is replaced by a quadratic majorizing cost $g_z(\mathbf{x}) \geq f_z(\mathbf{A}\mathbf{x})$ that coincides at the point $\mathbf{x} = \mathbf{x}^t$ (i.e., $g_z(\mathbf{x}^t) = f_z(\mathbf{A}\mathbf{x}^t)$), where majorization can be achieved via appropriate choice of $c > 0$. This approach is motivated by the fact that, if $f_x(\mathbf{x})$ and $f_z(\mathbf{z})$ are both separable, as in (2), then both the gradient in line 5 and minimization in line 6 can be performed componentwise. Moreover, when $f_x(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, as in the LASSO problem [41], the minimization in line 6 can be computed directly via a shrinkage and thresholding operation – hence the name of the algorithm. The convergence of the ISTA method tends to be slow, but a number of enhanced methods have been successful and widely-used [7]–[10].

C. Alternating Direction Method of Multipliers

A second common class of methods is built around the Alternating Direction Method of Multipliers (ADMM) [11] approach shown in Algorithm 3. The Lagrangian for the optimization problem (1) is given by

$$L(\mathbf{x}, \mathbf{z}, \mathbf{s}) := F(\mathbf{x}, \mathbf{z}) + \mathbf{s}^T(\mathbf{z} - \mathbf{A}\mathbf{x}), \quad (10)$$

where $\mathbf{s}$ are the dual parameters. ADMM attempts to produce a sequence of estimates $(\mathbf{x}^t, \mathbf{z}^t, \mathbf{s}^t)$ that converge to a saddle point of the Lagrangian (10). The parameters of the algorithm are a step-size $\alpha > 0$ and the penalty terms $Q_z(\cdot)$ and $Q_x(\cdot)$, which classical ADMM would choose as

$$Q_x(\mathbf{x}, \mathbf{x}^t, \mathbf{z}^t, \alpha) = \frac{\alpha}{2}\|\mathbf{z}^t - \mathbf{A}\mathbf{x}\|^2 \quad (11a)$$
$$Q_z(\mathbf{z}, \mathbf{z}^t, \mathbf{x}^{t+1}, \alpha) = \frac{\alpha}{2}\|\mathbf{z} - \mathbf{A}\mathbf{x}^{t+1}\|^2. \quad (11b)$$

Algorithm 3 Alternating Direction Method of Multipliers (ADMM)

Require: $\mathbf{A}$, α, functions $f_x(\cdot)$, $f_z(\cdot)$, $Q_x(\cdot)$, $Q_z(\cdot)$
1: $t \leftarrow 0$
2: Initialize $\mathbf{x}^t$, $\mathbf{z}^t$, $\mathbf{s}^t$
3: **repeat**
4: $\quad \mathbf{x}^{t+1} \leftarrow \arg\min_{\mathbf{x}} L(\mathbf{x}, \mathbf{z}^t, \mathbf{s}^t) + Q_x(\mathbf{x}, \mathbf{x}^t, \mathbf{z}^t, \alpha)$
5: $\quad \mathbf{z}^{t+1} \leftarrow \arg\min_{\mathbf{z}} L(\mathbf{x}^{t+1}, \mathbf{z}, \mathbf{s}^t) + Q_z(\mathbf{z}, \mathbf{z}^t, \mathbf{x}^{t+1}, \alpha)$
6: $\quad \mathbf{s}^{t+1} \leftarrow \mathbf{s}^t + \alpha(\mathbf{z}^{t+1} - \mathbf{A}\mathbf{x}^{t+1})$
7: **until** Terminated

When the objective function admits a separable form (2) and one uses the auxiliary function $Q_z(\cdot)$ in (11b), the $\mathbf{z}$-minimization in line 5 separates into m scalar optimizations. However, due to the quadratic term $\|\mathbf{A}\mathbf{x}\|^2$ in (11a), the $\mathbf{x}$-minimization in line 4 does not separate for general $\mathbf{A}$. To circumvent this problem, one might consider a separable inexact $\mathbf{x}$-minimization, since many inexact variants of ADMM are known to converge [12]. For example, $Q_x(\cdot)$ might be chosen to yield separability while majorizing the original cost in line 4, as was done for ISTA's line 6, i.e.,

$$Q_x(\mathbf{x}, \mathbf{x}^t, \mathbf{z}^t, \alpha) \quad (12)$$
$$= \frac{\alpha}{2}\|\mathbf{z}^t - \mathbf{A}\mathbf{x}\|^2 + \frac{1}{2}(\mathbf{x} - \mathbf{x}^t)^T(c\mathbf{I} - \alpha\mathbf{A}^T\mathbf{A})(\mathbf{x} - \mathbf{x}^t)$$

with appropriate c, after which ADMM's line 4 would become

$$\arg\min_{\mathbf{x}} f_x(\mathbf{x}) + \frac{c}{2}\left\|\mathbf{x} - \mathbf{x}^t + \frac{\alpha}{c}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^t - \mathbf{z}^t - \frac{1}{\alpha}\mathbf{s}^t\right)\right\|^2. \quad (13)$$

Many other choices of penalty $Q_x(\cdot)$ have also been considered in the literature (see, e.g., the overview in [18]), yielding connections to Douglas Rachford splitting [12], split Bregman [13], split inexact Uzawa [14], proximal forward-backward splitting [15], and various primal-dual algorithms [16]–[20].

Other variants of ADMM are also possible [11]. For example, the step-size α might vary with the iteration t, or the penalty terms might have the form $(\mathbf{z}-\mathbf{A}\mathbf{x})^T\mathbf{P}(\mathbf{z}-\mathbf{A}\mathbf{x})$ for positive semidefinite $\mathbf{P}$. As we will see, these generalizations provide a connection to GAMP.

III. Fixed-Points of Max-Sum GAMP

Our first result connects the max-sum GAMP algorithm to inexact ADMM. Given points $(\mathbf{x},\mathbf{z})$, define the matrices

$$\mathbf{Q}_x := \Big(\operatorname{diag}(\mathbf{d}_x)+\mathbf{A}^T\operatorname{diag}(\mathbf{d}_z)\mathbf{A}\Big)^{-1} \quad (14a)$$

$$\mathbf{Q}_z := \Big(\operatorname{diag}(\mathbf{d}_z)^{-1}+\mathbf{A}\operatorname{diag}(\mathbf{d}_x)^{-1}\mathbf{A}^T\Big)^{-1} \quad (14b)$$

where $\mathbf{d}_x$ and $\mathbf{d}_z$ are the componentwise second derivatives

$$\mathbf{d}_x := \frac{\partial^2 f_x(\mathbf{x})}{\partial \mathbf{x}^2}, \quad \mathbf{d}_z := \frac{\partial^2 f_z(\mathbf{z})}{\partial \mathbf{z}^2}. \quad (15)$$

Note that when $f_x(\mathbf{x})$ and $f_z(\mathbf{z})$ are strictly convex, the vectors $\mathbf{d}_x$ and $\mathbf{d}_z$ are positive. Observe that the matrix $\mathbf{Q}_x$ in (14) is the inverse Hessian of the objective function $F(\mathbf{x},\mathbf{z})$ constrained to $\mathbf{z}=\mathbf{A}\mathbf{x}$. That is,

$$\mathbf{Q}_x = \left[\frac{\partial^2}{\partial \mathbf{x}^2}F(\mathbf{x},\mathbf{A}\mathbf{x})\right]^{-1}.$$

Theorem 1: The outputs of the max-sum GAMP version of Algorithm 1 satisfy the recursions

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} L(\mathbf{x},\mathbf{z}^t,\mathbf{s}^t)+\frac{1}{2}\|\mathbf{x}-\mathbf{x}^t\|^2_{\boldsymbol{\tau}_r^t} \quad (16a)$$

$$\mathbf{z}^{t+1} = \arg\min_{\mathbf{z}} L(\mathbf{x}^{t+1},\mathbf{z},\mathbf{s}^t)+\frac{1}{2}\|\mathbf{z}-\mathbf{A}\mathbf{x}^{t+1}\|^2_{\boldsymbol{\tau}_p^{t+1}} \quad (16b)$$

$$\mathbf{s}^{t+1} = \mathbf{s}^t+\frac{1}{\boldsymbol{\tau}_p^{t+1}}(\mathbf{z}^{t+1}-\mathbf{A}\mathbf{x}^{t+1}) \quad (16c)$$

where $L(\mathbf{x},\mathbf{z},\mathbf{s})$ is the Lagrangian defined in (10).

Now suppose that $\widehat{\mathbf{x}}$, $\widehat{\mathbf{z}}$, $\mathbf{s}$, $\mathbf{p}$, $\boldsymbol{\tau}_r,\ldots$ are fixed points of the algorithm (where the "hats" on $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{z}}$ are used to distinguish them from dummy variables). Then, the fixed points are critical points of the constrained optimization (1) in that $\widehat{\mathbf{z}}=\mathbf{A}\widehat{\mathbf{x}}$ and

$$\frac{\partial}{\partial \mathbf{x}}L(\widehat{\mathbf{x}},\widehat{\mathbf{z}},\mathbf{s})=\mathbf{0}, \quad \frac{\partial}{\partial \mathbf{z}}L(\widehat{\mathbf{x}},\widehat{\mathbf{z}},\mathbf{s})=\mathbf{0}. \quad (17)$$

Also, the quadratic terms $\boldsymbol{\tau}_x,\boldsymbol{\tau}_s$ are the *approximate diagonals* (see Appendix A) to the inverse Hessian $\mathbf{Q}_x$ and the matrix $\mathbf{Q}_z$ in (14) at $(\mathbf{x},\mathbf{z})=(\widehat{\mathbf{x}},\widehat{\mathbf{z}})$.

Proof: See Appendix B. ∎

The first part of the Theorem, equation (16), shows that max-sum GAMP can be interpreted as the ADMM Algorithm 3 with adaptive vector-valued step-sizes $\boldsymbol{\tau}_r^t$ and $\boldsymbol{\tau}_p^t$ and a particular choice of penalty $Q_x(\cdot)$. To more precisely connect GAMP and existing algorithms, it helps to express GAMP's $\mathbf{x}$-update (16a) as the $\theta=0$ case of

$$\arg\min_{\mathbf{x}} f_x(\mathbf{x})+\frac{1}{2}\big\|\mathbf{x}-\mathbf{x}^t+\boldsymbol{\tau}_r^t\mathbf{A}^T\big(\theta(\mathbf{s}^{t-1}-\mathbf{s}^t)-\mathbf{s}^t\big)\big\|^2_{\boldsymbol{\tau}_r^t}, \quad (18)$$

and recognize that the ISTA-inspired inexact ADMM $\mathbf{x}$-update (13) coincides with the $\theta=1$ case under step-sizes $\alpha=1/\boldsymbol{\tau}_p^t$ and $c=1/\boldsymbol{\tau}_r^t$. The convergence of this algorithm for particular $\theta\in[0,1]$ was studied in [18]–[20] under convex functions $f_x(\cdot)$ and $f_z(\cdot)$ and *fixed scalar* step-sizes. Unfortunately, these convergence results do not directly apply to the adaptive vector-valued step-sizes of GAMP. However, the second part of the Theorem shows at least that, if the algorithm converges, its fixed points will be critical points of the constrained optimization (1). In addition, the quadratic term $\boldsymbol{\tau}_x$ can be interpreted as an approximate diagonal to the inverse Hessian.

Finally, it is useful to compare the fixed-points of GAMP with those of standard BP. A classic result of [30] shows that any fixed point for standard max-sum loopy BP is locally optimal in the sense that one cannot improve the objective function by perturbing the solution on any set of components whose variables belong to a subgraph that contains at most one cycle. In particular, if the overall graph is acyclic, any fixed-point of standard max-sum loopy BP is globally optimal. Also, for any graph, the objective function cannot be reduced by changing any individual component. The local optimality for GAMP provided by Theorem 1 is weaker in that the fixed-points only satisfy first-order conditions for saddle points of the Lagrangian. This implies that, even an individual component may only be locally optimal.

IV. Fixed-Points of Sum-Product GAMP

A. Variational Formulation

The characterization of the fixed points of the sum-product GAMP algorithm is somewhat more complicated to describe, and requires a certain *variational* interpretation – a common framework for understanding sum-product loopy BP [32], [34]. For any fixed observation $\mathbf{y}$, the density function $p_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\cdot,\cdot|\mathbf{y})$ in (3) must minimize

$$p_{\mathbf{x},\mathbf{z}|\mathbf{y}} = \arg\min_b D(b||p_{\mathbf{x},\mathbf{z}|\mathbf{y}}), \quad (19)$$

where the minimization is over all density functions $b(\mathbf{x},\mathbf{z})$ on the set $\mathbf{z}=\mathbf{A}\mathbf{x}$ and $D(b||p_{\mathbf{x},\mathbf{z}|\mathbf{y}})$ is the Kullback-Leibler (KL) divergence. Now, let $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ and $p_{\mathbf{z}|\mathbf{y}}(\mathbf{z}|\mathbf{y})$ be the marginal densities for the posterior $p_{\mathbf{x},\mathbf{z}|\mathbf{y}}$. Using the separable nature of $F(\mathbf{x},\mathbf{z})$ in (2), it can be verified that the minimization (19) can be rewritten as

$$(p_{\mathbf{x}|\mathbf{y}},p_{\mathbf{z}|\mathbf{y}}) = \arg\min_{b_x,b_z} J_{\mathrm{KL}}(b_x,b_z) \text{ s.t. } b_z=T_{\mathbf{A}}b_x, \quad (20)$$

where the minimization is over density functions $b_x(\mathbf{x})$ and $b_z(\mathbf{z})$ and $J_{\mathrm{KL}}(b_x,b_z)$ is the functional

$$J_{\mathrm{KL}}(b_x,b_z) := D(b_x||e^{-f_x})+D(b_z||e^{-f_z})+H(b_z). \quad (21)$$

In (20), we have used the notation $b_z=T_{\mathbf{A}}b_x$ to indicate that $b_z(\mathbf{z})$ is the density for a random vector $\mathbf{z}=\mathbf{A}\mathbf{x}$ with $\mathbf{x}\sim b_x(\mathbf{x})$. Thus, $p_{\mathbf{z}|\mathbf{y}}=T_{\mathbf{A}}p_{\mathbf{x}|\mathbf{y}}$. Note that we are treating $\mathbf{A}$ as deterministic.

Our next main result, Theorem 2 below, will show that the fixed points of the sum-product GAMP algorithm can be interpreted as critical points of the optimization (20), but with three key approximations: First, similar to what is known as a

mean-field approximation, the optimization is performed only over factorizable density functions of the form

$$b_x(\mathbf{x}) = \prod_{j=1}^{n} b_{x_j}(x_j), \quad b_z(\mathbf{z}) = \prod_{i=1}^{m} b_{z_i}(z_i). \tag{22}$$

Secondly, the objective function in (21) is replaced by

$$\begin{aligned} J_{\rm SP}(b_x, b_z, \boldsymbol{\tau}_p) &:= D(b_x||e^{-f_x}) + D(b_z||e^{-f_z}) \\ &\quad + H_{\rm gauss}(b_z, \boldsymbol{\tau}_p) \end{aligned} \tag{23}$$

where $\boldsymbol{\tau}^p$ is a positive vector and $H_{\rm gauss}(b_z, \boldsymbol{\tau}_p)$ is the Gaussian upper bound on the entropy

$$H_{\rm gauss}(b_z, \boldsymbol{\tau}_p) := \sum_{i=1}^{m} \left[\frac{1}{2\tau_{pi}} {\rm var}(z_i|b_{z_i}) + \frac{1}{2}\log(2\pi\tau_{pi}) \right]. \tag{24}$$

The third and final approximation is that the constraint that $b_z = T_{\mathbf{A}} b_x$ is replaced by weaker *moment matching* constraints $\mathbb{E}(\mathbf{z}|b_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|b_x)$ and $\boldsymbol{\tau}^p = \mathbf{S}{\rm var}(\mathbf{x}|b_x)$. The resulting optimization is

$$(\widehat{b}_x, \widehat{b}_z, \boldsymbol{\tau}_p) = \arg\min J_{\rm SP}(b_x, b_z, \boldsymbol{\tau}_p) \tag{25a}$$

$$\text{s.t.} \quad \mathbb{E}(\mathbf{z}|b_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|b_x), \quad \boldsymbol{\tau}^p = \mathbf{S}{\rm var}(\mathbf{x}|b_x). \tag{25b}$$

Note that in (25b), the variance ${\rm var}(\mathbf{x}|b_x)$ is to be interpreted as a vector (not a covariance matrix) with components ${\rm var}(x_j|b_{x_j})$. The next lemma provides a certain Gaussian interpretation to this approximate optimization (25).

Lemma 1: For any positive vector $\boldsymbol{\tau}_p$, and density functions b_x and b_z, $J_{\rm SP}(b_x, b_z, \boldsymbol{\tau}^p)$ is an upper bound:

$$J_{\rm SP}(b_x, b_z, \boldsymbol{\tau}_p) \geq J_{\rm KL}(b_x, b_z), \tag{26}$$

with equality in the case when b_z is separable and Gaussian and $\boldsymbol{\tau}_p = {\rm var}(\mathbf{z}|b_z)$.

Proof: See Appendix C. ∎

Thus, the optimization (25) can be interpreted as an approximation where the distributions are factorizable and the output distribution b_z is assumed to be Gaussian. We will thus call the optimization (25) the *Gaussian approximate optimization*. This Gaussian approximate optimization is consistent with the manner in which the sum-product GAMP algorithm is derived: In standard loopy belief propagation, the sum-product updates assume independent, and thus factorizable, distributions at the input and output nodes. Moreover, the GAMP variant of algorithm additionally applies a Central Limit Theorem approximation to justify that the output distributions are approximately Gaussian.

It is important to realize that the optimization (25) is neither necessarily an upper nor lower bound to (20): Although the objective function satisfies the upper bound (23), the moment matching constraints (25b) are weaker than $b_z = T_{\mathbf{A}} b_x$.

B. Equivalent Optimization

Corresponding to the variational objective function (23), define

$$F_{\rm SP}(\overline{\mathbf{x}}, \overline{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p) = F^x_{\rm SP}(\overline{\mathbf{x}}, \boldsymbol{\tau}_x) + F^z_{\rm SP}(\overline{\mathbf{z}}, \boldsymbol{\tau}_p). \tag{27}$$

where the terms on the right-hand side are the constrained optimizations

$$\begin{aligned} F^x_{\rm SP}(\overline{\mathbf{x}}, \boldsymbol{\tau}_x) &:= \min_{b_x} D(b_x||e^{-f_x}) \\ &\quad \text{s.t. } \mathbb{E}(\mathbf{x}|b_x) = \overline{\mathbf{x}}, \quad {\rm var}(\mathbf{x}|b_x) = \boldsymbol{\tau}_x \end{aligned} \tag{28a}$$

$$\begin{aligned} F^z_{\rm SP}(\overline{\mathbf{z}}, \boldsymbol{\tau}_p) &:= \min_{b_z} D(b_z||e^{-f_z}) + H_{\rm gauss}(b_z, \boldsymbol{\tau}_p) \\ &\quad \text{s.t. } \mathbb{E}(\mathbf{z}|b_z) = \overline{\mathbf{z}}. \end{aligned} \tag{28b}$$

Lemma 2: Suppose that $(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p)$ are solutions to the optimization

$$(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p) = \arg\min_{\overline{\mathbf{x}}, \overline{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p} F_{\rm SP}(\overline{\mathbf{x}}, \overline{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p) \tag{29a}$$

$$\text{s.t.} \quad \overline{\mathbf{z}} = \mathbf{A}\overline{\mathbf{x}}, \ \boldsymbol{\tau}_p = \mathbf{S}\boldsymbol{\tau}_x. \tag{29b}$$

Then the densities $(\widehat{b}_x, \widehat{b}_z)$ given by the solutions to the optimizations in (28) are minima of the variational optimization (25).

Conversely, given any solution $(\widehat{b}_x, \widehat{b}_z)$ to the variational optimization (25), the vectors

$$\widehat{\mathbf{x}} = \mathbb{E}(\mathbf{x}|\widehat{b}_x), \quad \boldsymbol{\tau}_x = {\rm var}(\mathbf{x}|\widehat{b}_x), \tag{30a}$$

$$\widehat{\mathbf{z}} = \mathbb{E}(\mathbf{z}|\widehat{b}_z), \quad \boldsymbol{\tau}_p = \mathbf{S}\boldsymbol{\tau}_x, \tag{30b}$$

are solutions to the optimization (29).

Proof: See Appendix D. ∎

The lemma shows that the variational optimization (29) over densities is equivalent to a vector-valued optimization (25). So, we need only consider the vector-valued optimization (25). Corresponding to this constrained optimization, define the Lagrangian

$$L_{\rm SP}(\overline{\mathbf{x}}, \overline{\mathbf{z}}, \boldsymbol{\tau}^x, \boldsymbol{\tau}^p, \mathbf{s}) = F_{\rm SP}(\overline{\mathbf{x}}, \overline{\mathbf{z}}, \boldsymbol{\tau}^x, \boldsymbol{\tau}^p) + \mathbf{s}^T(\overline{\mathbf{z}} - \mathbf{A}\overline{\mathbf{x}}), \tag{31}$$

where $\mathbf{s}$ represents a vector of dual parameters. We can now state the main result.

Theorem 2: Consider the outputs of the sum-product GAMP version of Algorithm 1. Then, the updates for $\mathbf{x}^t$ and $\boldsymbol{\tau}_x^t$ are equivalent to

$$\begin{aligned} (\mathbf{x}^{t+1}, \boldsymbol{\tau}_x^{t+1}) &= \arg\min_{\overline{\mathbf{x}}, \boldsymbol{\tau}_x} \Big[L_{\rm SP}(\overline{\mathbf{x}}, \mathbf{z}^t, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p^t, \mathbf{s}^t) \\ &\quad + \frac{1}{2}(\boldsymbol{\tau}_s^t)^T \mathbf{S}\boldsymbol{\tau}_x + \frac{1}{2}\|\overline{\mathbf{x}} - \mathbf{x}^t\|^2_{\boldsymbol{\tau}_r^t} \Big]. \end{aligned} \tag{32}$$

where $L_{\rm SP}(\mathbf{x}, \mathbf{z}, \mathbf{s})$ is the Lagrangian in (10). In addition, the updates for $\mathbf{z}^t$, $\boldsymbol{\tau}_p^t$ and $\mathbf{s}^t$ are equivalent to

$$\boldsymbol{\tau}_p^{t+1} = \mathbf{S}\boldsymbol{\tau}_x^{t+1} \tag{33a}$$

$$\begin{aligned} \mathbf{z}^{t+1} &= \arg\min_{\overline{\mathbf{z}}} \Big[L_{\rm SP}(\mathbf{x}^{t+1}, \overline{\mathbf{z}}, \boldsymbol{\tau}_x^{t+1}, \boldsymbol{\tau}_p^{t+1}, \mathbf{s}^t) \\ &\quad + \frac{1}{2}\|\overline{\mathbf{z}} - \mathbf{A}\mathbf{x}^{t+1}\|^2_{\boldsymbol{\tau}_p^{t+1}} \Big] \end{aligned} \tag{33b}$$

$$\mathbf{s}^{t+1} = \mathbf{s}^t + \frac{1}{\boldsymbol{\tau}_p^{t+1}}(\mathbf{z}^{t+1} - \mathbf{A}\mathbf{x}^{t+1}) \tag{33c}$$

$$\boldsymbol{\tau}_s^{t+1} = 2\frac{\partial}{\partial \boldsymbol{\tau}^p} L_{\rm SP}(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}, \boldsymbol{\tau}_x^{t+1}, \boldsymbol{\tau}_p^{t+1}, \mathbf{s}^{t+1}). \tag{33d}$$

Moreover, any fixed point of the sum-product GAMP algorithm is a critical point of the Lagrangian (31). In addition,

the density functions for the minimization in (28) are given by

$$\widehat{b}^x(\mathbf{x}) = p(\mathbf{x}|\mathbf{r},\boldsymbol{\tau}^r), \quad \widehat{b}^z(\mathbf{z}) = p(\mathbf{z}|\mathbf{p},\boldsymbol{\tau}^p), \tag{34}$$

where $p(\mathbf{x}|\mathbf{r},\boldsymbol{\tau}^r)$ and $p(\mathbf{z}|\mathbf{p},\boldsymbol{\tau}^p)$ are given by (8).

Proof: See Appendix E. ∎

Theorem 2 shows a relation between sum-product GAMP and both the ISTA and ADMM methods described earlier. Specifically, define the variables

$$\mathbf{u} := (\overline{\mathbf{x}}, \boldsymbol{\tau}_x), \quad \mathbf{v} := (\overline{\mathbf{z}}, \boldsymbol{\tau}_p).$$

Due to the separable structure of the objective function (28), the optimization (29) can be regarded as minimizing a separable function $F^x_{\rm SP}(\mathbf{u}) + F^z_{\rm SP}(\mathbf{v})$ with linear constraints (29b) between $\mathbf{u}$ and $\mathbf{v}$. In this context, the $\mathbf{x}$ and $\mathbf{z}$ minimizations in (32) and (33b) follow the format of the ADMM minimizations in Algorithm 3 for certain choices of the auxiliary functions. On the other hand, the optimization over $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_p$ components follow the gradient-based method in the generalized ISTA method in Algorithm 2. So, the sum-product GAMP algorithm can be seen as a hybrid of the ISTA and ADMM methods for the optimization (29), which is equivalent to the variational optimization (25).

Unfortunately, this hybrid ISTA-ADMM method is non-standard and there is no existing convergence theory on the algorithm. However, Theorem 2 at least shows that if the sum-product GAMP algorithm converges, its fixed points correspond to critical points of optimization (29).

It is useful to briefly compare Theorem 2 with the variational interpretation of standard loopy BP. It is well-known [32] that the fixed points of standard loopy BP can be interpreted as distributions on the factor and variable nodes that minimize the so-called Bethe free energy subject to certain local consistency constraints. In comparison, GAMP appears to minimize a Gaussian approximation of the KL divergence subject to weaker moment matching constraints between the distributions on the variable nodes. In this manner, the fixed-points of GAMP appears closer in form of those of expectation propagation (EP) methods that can also be interpreted as saddle points of a certain free energy subject to moment matching [33]. However, the exact relation between EP and sum-product GAMP fixed points requires further study.

Conclusions

Although AMP methods admit precise analyses in the context of large random transform matrices $\mathbf{A}$, their behavior for general matrices is less well-understood. This limitation is unfortunate since many transforms arising in practical problems such as imaging and regression are not well-modeled as realizations of large random matrices. To help overcome these limitations, this paper draws connections between AMP and certain variants of standard optimization methods that employ adaptive vector-valued step-sizes. These connections enable a precise characterization of the fixed-points of both max-sum and sum-product GAMP for the case of arbitrary transform matrices $\mathbf{A}$. The convergence of AMP methods for general $\mathbf{A}$ is, however, still not fully understood. Simulations (not shown here) have indicated, for example, that under general choices of $\mathbf{A}$, AMP may diverge. We hope that the connections between AMP and standard optimization methods provided here help to better understand, and even improve, AMP convergence with general matrices.

Appendix A
Approximate Diagonals

Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and positive vectors $\mathbf{d}_x \in \mathbb{R}^n$ and $\mathbf{d}_z$, consider the positive matrices (14). We analyze these asymptotic behavior of these matrices under the following assumptions:

Assumption 1: Consider a sequence of matrices $\mathbf{Q}_x$ and $\mathbf{Q}_z$ of the form (14), indexed by the dimension n satisfying:

(a) The dimension m is a deterministic function of n with $\lim_{n\to\infty} m/n = \beta$ for some $\beta > 0$,

(b) The positive vectors $\mathbf{d}_x$ and $\mathbf{d}_z$ are deterministic vectors with

$$\limsup_{n\to\infty} \|\mathbf{d}_x\|_\infty < \infty, \quad \limsup_{n\to\infty} \|\mathbf{d}_z\|_\infty < \infty.$$

(c) The components of $\mathbf{A}$ are independent, zero-mean with $\text{var}(A_{ij}) = S_{ij}$ for some deterministic matrix $\mathbf{S}$ such that

$$\limsup_{n} \max_{i,j} nS_{ij} < \infty.$$

Theorem 3 ([42]): Consider a sequence of matrices $\mathbf{Q}_x$ and $\mathbf{Q}_z$ in Assumption 1. Then, for each n, there exists positive vectors let $\boldsymbol{\xi}_x$ and $\boldsymbol{\xi}_z$ nonlinear equations

$$\frac{1}{\boldsymbol{\xi}_z} = \frac{1}{\mathbf{d}_z} + \mathbf{S}\boldsymbol{\xi}_x, \quad \frac{1}{\boldsymbol{\xi}_x} = \mathbf{d}_x + \mathbf{S}^T\boldsymbol{\xi}_z, \tag{35}$$

where the inverses on the left side are componentwise. Moreover, the vectors $\boldsymbol{\xi}_z$ and $\boldsymbol{\xi}_x$ are asymptotic diagonals of $\mathbf{Q}_x$ and $\mathbf{Q}_z$ in the following sense: For any deterministic sequence of positive vectors $\mathbf{u}_x \in \mathbb{R}^n$ and $\mathbf{u}_z \in \mathbb{R}^m$, such that

$$\limsup_{n\to\infty} \|\mathbf{u}_x\|_\infty < \infty, \quad \limsup_{n\to\infty} \|\mathbf{u}_z\|_\infty < \infty,$$

the following limits hold almost surely

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^n [u_{xj}((Q_x)_{jj} - \xi_{xj})] &= 0\\ \lim_{n\to\infty} \frac{1}{m}\sum_{i=1}^m [u_{zi}((Q_z)_{ii} - \xi_{zi})] &= 0.\end{aligned}$$

Proof: This result is a special case of the results in [42]. ∎

The results says that, for certain large random matrices $\mathbf{A}$, $\boldsymbol{\xi}_x$ and $\boldsymbol{\xi}_z$ are approximate diagonals of the matrices $\mathbf{Q}_x$ and $\mathbf{Q}_z$, respectively. This motivates the following definition for deterministic $\mathbf{A}$.

Definition 1: Consider matrices $\mathbf{Q}_x$ and $\mathbf{Q}_z$ of the form (14) for some *deterministic* (i.e. non-random) $\mathbf{A}$, $\mathbf{d}_x$ and $\mathbf{d}_z$. Let $\mathbf{S} = |\mathbf{A}|^2$ be the componentwise magnitude squared of $\mathbf{A}$. Then, the unique positive solutions $\boldsymbol{\xi}_z$ and $\boldsymbol{\xi}_x$ to (35) will be called the *approximate diagonals* of $\mathbf{Q}_x$ and $\mathbf{Q}_z$.

APPENDIX B
PROOF OF THEOREM 1

To prove (16b), observe that

$$\begin{aligned}
&\arg\max_{\mathbf{z}}\Big[L(\mathbf{x}^t,\mathbf{z},\mathbf{s}^{t-1})+\frac{1}{2}\|\mathbf{z}-\mathbf{A}\mathbf{x}^t\|^2_{\boldsymbol{\tau}_p^t}\Big]\\
&\overset{(a)}{=}\ \arg\max_{\mathbf{z}}\Big[f_z(\mathbf{z})+(\mathbf{s}^{t-1})^T\mathbf{z}+\frac{1}{2}\|\overline{\mathbf{z}}-\mathbf{A}\mathbf{x}^t\|^2_{\boldsymbol{\tau}_p^t}\Big]\\
&\overset{(b)}{=}\ \arg\max_{\mathbf{z}} F_z(\mathbf{z},\mathbf{p}^t,\boldsymbol{\tau}_p^t)\overset{(c)}{=}\mathbf{z}^t,
\end{aligned}$$

where (a) follows from substituting (2) and (10) into (16b) and eliminating the terms that do not depend on $\mathbf{z}$ and (b) follows from the definition of $\mathbf{p}^t$ in line 8 and F_z in (4b); and (c) is the definition of $\mathbf{z}^t$ in line 10. This proves (16b). The update (16a) is proven similarly. To prove (16c), observe that

$$\mathbf{s}^t\overset{(a)}{=}\frac{1}{\boldsymbol{\tau}_p^t}(\mathbf{z}^t-\mathbf{p}^t)\overset{(b)}{=}\mathbf{s}^{t-1}+\frac{1}{\boldsymbol{\tau}_p^t}(\mathbf{z}^t-\mathbf{A}\mathbf{x}^t)$$

where (a) follows from the update of $\mathbf{s}^t$ in line 16 in Algorithm 1 (recall that the division is componentwise); and (b) follows from the update for $\mathbf{p}^t$ in line 8. We have thus proven the equivalence of the max-sum GAMP algorithm with the Lagrangian updates (16).

Now consider any fixed point of the max-sum GAMP algorithm. A fixed point of (16c), requires that

$$\widehat{\mathbf{z}}=\mathbf{A}\widehat{\mathbf{x}} \tag{36}$$

so the vector satisfies the constraint of the optimization (1). Now, using (36) and the fact that $\widehat{\mathbf{z}}$ is the maxima of (16b), we have that

$$\frac{\partial}{\partial\mathbf{z}}L(\widehat{\mathbf{x}},\widehat{\mathbf{z}},\mathbf{s})=0.$$

Similarly, since $\mathbf{x}$ is the maxima of (16a), we have that

$$\frac{\partial}{\partial\mathbf{x}}L_{\rm MS}(\widehat{\mathbf{x}},\widehat{\mathbf{z}},\mathbf{s})=0.$$

Thus, the fixed point $(\widehat{\mathbf{x}},\widehat{\mathbf{z}},\mathbf{s})$ is a critical point of the Lagrangian (10).

Finally, consider the quadratic terms $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_r$ at a fixed point. From the updates of the $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_r$ in Algorithm 1 and the definition of $\mathbf{d}_x$ in (15), we obtain

$$\frac{1}{\boldsymbol{\tau}_x}=\mathbf{d}_x+\frac{1}{\boldsymbol{\tau}_r}=\mathbf{d}_x+\mathbf{S}^T\boldsymbol{\tau}_s. \tag{37}$$

Similarly, the updates of $\boldsymbol{\tau}_s$ and $\boldsymbol{\tau}_p$ show that

$$\frac{1}{\boldsymbol{\tau}_s}=\frac{1}{\mathbf{d}_z}+\boldsymbol{\tau}_p=\frac{1}{\mathbf{d}_z}+\mathbf{S}\boldsymbol{\tau}_x. \tag{38}$$

From Definition 1, $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_s$ are approximate diagonals of $\mathbf{Q}_x$ and $\mathbf{Q}_z$ in (14).

APPENDIX C
PROOF OF LEMMA 1

For any positive vector $\boldsymbol{\tau}_p$ and density function b_z (even if it is not separable), we have the bound

$$\begin{aligned}
H(b_z)&\overset{(a)}{\le}\sum_{i=1}^m H(b_{z_i})\\
&\overset{(b)}{\le}\ \frac{1}{2}\sum_{i=1}^m\log(2\pi e\,\mathrm{var}(z_i|b_{z_i}))\\
&\overset{(c)}{\le}\ \frac{1}{2}\sum_{i=1}^m\left[\frac{\mathrm{var}(z_i|b_{z_i})}{\tau_{p_i}}+\log(2\pi\tau_{p_i})\right]\\
&\overset{(d)}{=}\ H_{\rm gauss}(b_z,\boldsymbol{\tau}_p)
\end{aligned} \tag{39}$$

where in (a), b_{z_i} is the marginal distribution on the component z_i of $\mathbf{z}$; (b) is the Gaussian upper bound on the entropy of each marginal distribution, (c) uses the fact that

$$\log(2\pi ev)\le\frac{v}{\tau}+\log(2\pi\tau),$$

for all τ with equality when $\tau=v$ and (d) is the definition of $H_{\rm gauss}$ in (24). When b_z is separable and Gaussian all the inequalities are equalities in (39). The inequality (26) now follows from comparing (21) and (23).

APPENDIX D
PROOF OF LEMMA 2

It is useful here and in subsequent proofs to introduce the following notation. Partition the objective function (23) as

$$J_{\rm SP}(b_x,b_z,\boldsymbol{\tau}_p)=J^x_{\rm SP}(b_x)+J^z_{\rm SP}(b_z,\boldsymbol{\tau}_p), \tag{40}$$

where

$$\begin{aligned}
J^x_{\rm SP}(b_x)&:=\ D(b_x||e^{-f_x}) &(41a)\\
J^z_{\rm SP}(b_z,\boldsymbol{\tau}_p)&:=\ D(b_z||e^{-f_z})+H_{\rm gauss}(b_z,\boldsymbol{\tau}_p). &(41b)
\end{aligned}$$

Then, we can rewrite (28) as

$$\begin{aligned}
&F^x_{\rm SP}(\overline{\mathbf{x}},\boldsymbol{\tau}_x):=\min_{b_x}J^x_{\rm SP}(b_x)\\
&\qquad\text{s.t. }\mathbb{E}(\mathbf{x}|b_x)=\overline{\mathbf{x}},\quad \mathrm{var}(\mathbf{x}|b_x)=\boldsymbol{\tau}_x &(42a)\\
&F^z_{\rm SP}(\overline{\mathbf{z}},\boldsymbol{\tau}_p):=\min_{b_z}J^z_{\rm SP}(b_z,\boldsymbol{\tau}_p)\\
&\qquad\text{s.t. }\mathbb{E}(\mathbf{z}|b_z)=\overline{\mathbf{z}}. &(42b)
\end{aligned}$$

Now, fix a positive vector $\boldsymbol{\tau}_p$ and consider the minimization (25) with the additional constraints that

$$\begin{aligned}
\overline{\mathbf{x}}&=\ \mathbb{E}(\mathbf{x}|\widehat{b}_x),\quad \boldsymbol{\tau}_x=\mathrm{var}(\mathbf{x}|\widehat{b}_x), &(43a)\\
\overline{\mathbf{z}}&=\ \mathbb{E}(\mathbf{z}|\widehat{b}^z) &(43b)
\end{aligned}$$

for some vectors $\overline{\mathbf{x}}$, $\overline{\mathbf{z}}$, $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_p$. Then, using the partition (40), the minima of (25) under the constraints (43) is precisely the function $F_{\rm SP}(\overline{\mathbf{x}},\overline{\mathbf{z}},\boldsymbol{\tau}_x,\boldsymbol{\tau}_p)$ in (27). Thus, the minimization (25) can be achieved by minimizing $F_{\rm SP}(\overline{\mathbf{x}},\overline{\mathbf{z}},\boldsymbol{\tau}_x,\boldsymbol{\tau}_p)$ under the constraints that $\overline{\mathbf{z}}=\mathbf{A}\overline{\mathbf{x}}$ and $\boldsymbol{\tau}_p=\mathbf{S}\boldsymbol{\tau}_x$. This minimization is precisely the optimization (29) and this fact establishes the equivalence between the two optimizations.

APPENDIX E
PROOF OF THEOREM 2

Our proof will now follow three parts: First, we will provide an explicit characterization for optimization problems of the form (28). Next, we will use this characterization to prove that the sum-product GAMP updates are equivalent to the ADMM-ISTA iterations in (32) and (33). Finally, we will show that the fixed points of the iterations correspond to critical points of the Lagrangian (31).

A. KL Minimization with Moment Constraints

We begin with a standard result on the minimization of the KL divergence subject to moment constraints.

Lemma 3: Let $f(u)$ be any real-valued measurable function on a real variable u. Given a mean value $\overline{u}$ and variance $\tau_u > 0$, then the following are equivalent statements about a probability density function $\widehat{b}(u)$:

(a) The probability density function $\widehat{b}(u)$ is the solution to the constrained optimization:

$$\widehat{b} = \arg\min_b D(b||e^{-f}) \quad \text{s.t. } \mathbb{E}(u|b) = \overline{u},\ \text{var}(u|b) = \tau_u. \tag{44}$$

(b) There exists q and $\tau_q > 0$ such that the density function is the solution to the unconstrained optimization

$$\widehat{b} = \arg\min_b \Big[D(b||e^{-f}) + \frac{1}{2\tau_q}\mathbb{E}((u-q)^2|b)\Big], \tag{45}$$

and $\mathbb{E}(u|\widehat{b}) = \overline{u}$, $\text{var}(u|\widehat{b}) = \tau_u$.

(c) There exists q and $\tau_q > 0$ such that the density function is of the form

$$\widehat{b} \propto \exp\left[-f(u) - \frac{1}{2\tau_q}(u-q)^2 \right], \tag{46}$$

and $\mathbb{E}(u|b) = \overline{u}$, $\text{var}(u|\widehat{b}) = \tau_u$.

Proof: This result is standard – similar calculations are in [34]. The equivalence between (a) and (b) can be shown via a Lagrangian argument and the equivalence between (b) and (c) can be found by taking the derivatives of the unconstrained objective (45) with respect to $b(u)$ for each u. ∎

B. Equivalence of GAMP and ADMM-ISTA Iterations

We now use Lemma 3 to that the sum-product GAMP iterations are equivalent to the ADMM-ISTA iterations in Theorem 2. We begin by proving (33b). Let $\tilde{\mathbf{z}}^t$ equal the right-hand side of (33b). We want to show that $\tilde{\mathbf{z}}^t = \mathbf{z}^t$, where $\mathbf{z}^t$ is the output of line 13 of the sum-product GAMP algorithm. To show this, we first observe that

$$\begin{aligned}\tilde{\mathbf{z}}^t &\overset{(a)}{=} \arg\min_{\overline{\mathbf{z}}}\Big[F^z_{\rm SP}(\overline{\mathbf{z}},\boldsymbol{\tau}_p^t) + (\mathbf{s}^{t-1})^T\overline{\mathbf{z}} + \frac{1}{2}\|\overline{\mathbf{z}} - \mathbf{A}\mathbf{x}^t\|^2_{\boldsymbol{\tau}_p^t}\Big] \\ &\overset{(b)}{=} \arg\min_{\overline{\mathbf{z}}}\Big[F^z_{\rm SP}(\overline{\mathbf{z}},\boldsymbol{\tau}_p^t) + \frac{1}{2}\|\overline{\mathbf{z}} - \mathbf{p}^t\|^2_{\boldsymbol{\tau}_p^t}\Big],\end{aligned} \tag{47}$$

where (a) follows from substituting (42b), (27) and (31) into (33b) and eliminating the terms that do not depend on $\overline{\mathbf{z}}$ and (b) follows from substituting in the definition of $\mathbf{p}^t$ in line (8). Now, using the definition of $F^z_{\rm SP}$ in (42b), it follows from (47) that

$$\tilde{\mathbf{z}}^t = \mathbb{E}[\mathbf{z}|\widehat{b}_z], \tag{48}$$

where $\widehat{b}_z$ is the density function on $\mathbf{z}$ that minimizes

$$\widehat{b}_z = \arg\min_{b_z}\Big[J^z_{\rm SP}(b_z,\boldsymbol{\tau}_p^t) + \frac{1}{2}\left\|\mathbb{E}(\mathbf{z}|b_z) - \mathbf{p}^t\right\|^2_{\boldsymbol{\tau}_p^t}\Big]. \tag{49}$$

Now, this minimization can be simplified as

$$\begin{aligned}\widehat{b}_z &\overset{(a)}{=} \arg\min_{b_z}\Big[D(b_z||e^{-f_z}) + \sum_{i=1}^m \frac{\text{var}(z_i|b_{z_i})}{2\tau^t_{p_i}} \\ &\qquad + \frac{1}{2}\left\|\mathbb{E}(\mathbf{z}|b_z) - \mathbf{p}^t\right\|^2_{\boldsymbol{\tau}_p^t}\Big] \\ &= \arg\min_{b_z}\sum_{i=1}^m\Big[D(b_{z_i}||e^{-f_{z_i}}) + \frac{\mathbb{E}(z_i - p_i^t)^2}{2\tau^t_{p_i}}\Big]\end{aligned} \tag{50}$$

where (a) follows from substituting (41b) and (24) into (49) and removing terms that do not depend on b_z; and (b) follows from the separability assumption (22) and the fact that, for any density function b_{z_i},

$$\text{var}(z_i|b_{z_i}) + \left|\mathbb{E}(z_i|b_{z_i}) - p_i^t\right|^2 = \mathbb{E}\left(|z_i - p_i^t|^2|b_{z_i}\right).$$

The minimization (49) is separable with solution

$$\widehat{b}_z(\mathbf{z}) = \prod_{i=1}^m \widehat{b}_{z_i}(z_i), \tag{51}$$

where the marginal density functions are the solutions to

$$\widehat{b}_{z_i}(z_i) = \arg\min_{b_{z_i}}\Big[D(b_{z_i}||e^{-f_{z_i}}) + \frac{\mathbb{E}(z_i - p_i^t)^2}{2\tau^t_{p_i}}\Big]. \tag{52}$$

From Lemma 3, the solution to the minimization (52) is given by

$$\widehat{b}_{z_i}(z_i) \propto \exp\left[f_{z_i}(z_i) - \frac{(z_i - p_i^t)^2}{2\tau^t_{p_i}}\right]. \tag{53}$$

Comparing (51) and (52) with (9) we see that

$$\widehat{b}_z(\mathbf{z}) = p(\mathbf{z}|\mathbf{p}^t,\boldsymbol{\tau}_p^t). \tag{54}$$

Hence, the expectation in (48) is precisely

$$\tilde{\mathbf{z}}^t = \mathbb{E}(\mathbf{z}|\mathbf{p}^t,\boldsymbol{\tau}_p^t),$$

which agrees with the definition of $\mathbf{z}^t$ in line 13 of Algorithm 1. Therefore, $\mathbf{z}^t = \tilde{\mathbf{z}}^t$ and we have proven (33b).

The proof of (32) is similar and also shows that $\widehat{\mathbf{x}}^{t+1}$ and $\boldsymbol{\tau}_x^{t+1}$ are the mean and variance of

$$\widehat{b}_x(\mathbf{x}) = p(\mathbf{x}|\mathbf{r}^t,\boldsymbol{\tau}_r^t). \tag{55}$$

The proof of (33c) is identical to the proof of (16c).

Finally to prove (33d), we take the derivatives

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\tau}_p}L_{\rm SP}(\widehat{\mathbf{x}}^t,\mathbf{z}^t,\boldsymbol{\tau}_x^t,\boldsymbol{\tau}_p^t,\mathbf{s}^t) \\ &\overset{(a)}{=} \frac{\partial}{\partial\boldsymbol{\tau}_p}F^z_{\rm SP}(\mathbf{z}^t,\boldsymbol{\tau}_p^t) \\ &\overset{(b)}{=} \frac{\partial}{\partial\boldsymbol{\tau}_p}J^z_{\rm SP}(\widehat{b}_z,\boldsymbol{\tau}_p^t) \\ &\overset{(c)}{=} \frac{\partial}{\partial\boldsymbol{\tau}_p}H_{\rm gauss}(\widehat{b}_z,\boldsymbol{\tau}_p^t) \\ &\overset{(d)}{=} \frac{1}{2}\Big[\frac{1}{\boldsymbol{\tau}_p^t} - \frac{\boldsymbol{\tau}_z^t}{(\boldsymbol{\tau}_p^t)^2}\Big] \overset{(e)}{=} \frac{1}{2}\boldsymbol{\tau}_s^t\end{aligned}$$

where (a) follows from substituting in (27) and (31) and removing the terms that do not depend on $\boldsymbol{\tau}_p$; (b) follows from the definition of $F^z_{\rm SP}$ in (42b); (c) follows from the definition of $J^z_{\rm SP}$ in (41b); (d) can be verified by simply taking the derivative of $H_{\rm gauss}$ with respect to each component τ_{p_i}, and (e) follows from the definition of $\boldsymbol{\tau}_s^t$ in line 17 of Algorithm 1. This proves (33d), and we have established that the sum-product GAMP updates are equivalent to (32) and (33).

C. Characterization of the Fixed Points

We conclude by showing that the fixed points of the sum-product GAMP algorithm are critical points of the Lagrangian (31). To account for the constraint that $\boldsymbol{\tau}_p = \mathbf{S}\boldsymbol{\tau}_x$, define the *modified Lagrangian*,

$$L_{\rm SP-mod}(\overline{\mathbf{x}}, \overline{\mathbf{z}}, \boldsymbol{\tau}_x, \mathbf{s}) = L_{\rm SP}(\overline{\mathbf{x}}, \overline{\mathbf{z}}, \boldsymbol{\tau}_x, \mathbf{S}\boldsymbol{\tau}_x, \mathbf{s}), \tag{56}$$

which is the Lagrangian $L_{\rm SP}(\overline{\mathbf{x}}, \overline{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p, \mathbf{s})$, with $\boldsymbol{\tau}_p = \mathbf{S}\boldsymbol{\tau}_x$.

Now consider any fixed point of the sum-product GAMP updates. Since they are fixed points, we will drop the dependence on t, and write the fixed points as $\widehat{\mathbf{x}}$, $\widehat{\mathbf{z}}$, $\mathbf{s}$, etc. To show that the fixed points are critical points of the optimization (29), we will show that the vectors are critical points of the modified Lagrangian $L_{\rm SP-mod}$ and satisfy the constraint that $\widehat{\mathbf{z}} = \mathbf{A}\widehat{\mathbf{x}}$.

Now, the fixed points of the sum-product GAMP algorithm must be fixed points of the Lagrangian updates (33) and (32). In particular, a fixed point of (33c), requires that

$$\widehat{\mathbf{z}} = \mathbf{A}\widehat{\mathbf{x}}. \tag{57}$$

Also, from (33a), we have that

$$\boldsymbol{\tau}_p = \mathbf{S}\boldsymbol{\tau}_x, \tag{58}$$

so the vectors satisfies the constraints (29b) in the optimization.

Now, using (57) and the fact that $\widehat{\mathbf{z}}$ is the minima of (33b), we have that

$$\frac{\partial}{\partial \overline{\mathbf{z}}} L_{\rm SP}(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p, \mathbf{s}) = 0. \tag{59}$$

Due to (58), the derivative (59) implies that

$$\frac{\partial}{\partial \overline{\mathbf{z}}} L_{\rm SP-mod}(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \mathbf{s}) = 0. \tag{60}$$

Similarly, since $\widehat{\mathbf{x}}$ is the minima of (32), we have that

$$\frac{\partial}{\partial \overline{\mathbf{x}}} L_{\rm SP-mod}(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \mathbf{s}) = 0. \tag{61}$$

The minimization (32) also implies that

$$\frac{\partial}{\partial \boldsymbol{\tau}_x} L_{\rm SP}(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p \mathbf{s}) = -\frac{1}{2}\mathbf{S}^T \boldsymbol{\tau}_s. \tag{62}$$

Therefore,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\tau}_x} L_{\rm SP-mod}(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \mathbf{s}) \\
&\overset{(a)}{=} \frac{\partial}{\partial \boldsymbol{\tau}_x} L_{\rm SP}(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p, \mathbf{s}) \\
&\quad + \mathbf{S}^T \frac{\partial}{\partial \boldsymbol{\tau}_p} L_{\rm SP}(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_p, \mathbf{s}) \\
&\overset{(b)}{=} -\frac{1}{2}\mathbf{S}^T \boldsymbol{\tau}_s + + \frac{1}{2}\mathbf{S}^T \boldsymbol{\tau}_s = 0,
\end{aligned} \tag{63}$$

where (a) follows from the the definition $L_{\rm SP-mod}$ in (56); (b) follows from (33d) and (62). The derivatives (61), (60) and (61) along with the constraints (57) and (58) show that the vectors $\widehat{\mathbf{x}}$, $\widehat{\mathbf{z}}$, $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_p$ are critical points of the optimization (29). Finally, using Lemma 3 and similar arguments as the derivation of (54) and (55) show that the density functions are $\widehat{b}_x$ and $\widehat{b}_z$ in the minimizations in (28) are given by (34).

REFERENCES

[1] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *J. Royal Stat. Soc. Series A*, vol. 135, pp. 370–385, 1972.

[2] P. McCullagh and J. A. Nelder, *Generalized linear models*, 2nd ed. Chapman & Hall, 1989.

[3] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. New York: Cambridge Univ. Press, 2012.

[4] A. Chambolle, R. A. DeVore, N. Y. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 319–335, Mar. 1998.

[5] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004.

[6] C. Vonesch and M. Unser, "Fast iterative thresholding algorithm for wavelet-regularized deconvolution," in *Proc. SPIE: Wavelet XII*, San Diego, CA, 2012.

[7] S. J. Wright, R. D. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.

[8] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problem," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[9] Y. E. Nesterov, "Gradient methods for minimizing composite objective function," *CORE Report*, 2007.

[10] J. Bioucas-Dias and M. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992 – 3004, Dec. 2007.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, 2010.

[12] J. Eckstein and D. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 5, pp. 293–318, 1992.

[13] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM J. Imaging Sciences*, vol. 2, no. 2, 2009.

[14] X. Zhang, M. Burger, and S. Osher, "A unified primal-dual algorithm framework based on Bregman iteration," *SIAM J. Sci. Comput.*, vol. 46, pp. 20–46, 2011.

[15] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, pp. 1168–1200, 2005.

[16] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM J. Control and Optimization*, vol. 29, no. 1, pp. 119–138, Jan. 1991.

[17] M. Zhu and T. Chan, "An efficient primal-dual hybrid gradient algorithm for total variation image restoration," UCLA CAM, Tech. Rep. 08-34, 2008.

[18] J. E. Esser, "Primal dual algorithms for convex models and applications to image restoration, registration and nonlocal inpainting," Ph.D. dissertation, The University of California, Los Angeles, 2010.

[19] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.*, vol. 40, pp. 120–145, 2011.

[20] B. He and X. Yuan, "Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective," *SIAM J. Imaging Sci.*, vol. 5, no. 1, pp. 119–149, 2012.

[21] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[22] ——, "Message passing algorithms for compressed sensing I: motivation and construction," in *Proc. Info. Theory Workshop*, Jan. 2010.

[23] ——, "Message passing algorithms for compressed sensing II: analysis and validation," in *Proc. Info. Theory Workshop*, Jan. 2010.

[24] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[25] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," in *Proc. Conf. on Inform. Sci. & Sys.*, Princeton, NJ, Mar. 2010, pp. 1–6.

[26] ——, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inform. Theory*, Saint Petersburg, Russia, Jul.–Aug. 2011, pp. 2174–2178.

[27] G. Caire, S. Shamai, A. Tulino, and S. Verdú, "Support recovery in compressed sensing: Information-theoretic bounds," in *Proc. UCSD Workshop Inform. Theory & Its Applications*, La Jolla, CA, Jan. 2011.

[28] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and iterative algorithms," in *Proc. ISIT*, Jul. 2012, pp. 1643 –1647.

[29] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," arXiv:1211.5164 [math.PR]., Nov. 2012.

[30] Y. Weiss and W. T. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 736–744, Feb. 2001.

[31] T. Heskes, "Stable fixed points of loopy belief propagation are minima of the Bethe free energy," in *Proc. Neural Information Process. Syst.*, Vancouver, Canada, Dec. 2003.

[32] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," in *Exploring Artificial Intelligence in the New Millennium*. San Francisco, CA: Morgan Kaufmann Publishers, 2003, pp. 239–269.

[33] M. Opper and O. Winther, "Expectation consistent approximate inference," *J. Mach. Learning Res.*, vol. 1, pp. 2177–2204, 2005.

[34] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, 2008.

[35] J. Boutros and G. Caire, "Iterative multiuser joint decoding: Unified framework and asymptotic analysis," *IEEE Trans. Inform. Theory*, vol. 48, no. 7, pp. 1772–1793, Jul. 2002.

[36] T. Tanaka and M. Okada, "Approximate belief propagation, density evolution, and neurodynamics for CDMA multiuser detection," *IEEE Trans. Inform. Theory*, vol. 51, no. 2, pp. 700–706, Feb. 2005.

[37] D. Guo and C.-C. Wang, "Asymptotic mean-square optimality of belief propagation for sparse linear systems," in *Proc. IEEE Inform. Theory Workshop*, Chengdu, China, Oct. 2006, pp. 194–198.

[38] A. Montanari, "Graphical model concepts in compressed sensing," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge Univ. Press, Jun. 2012, pp. 394–438.

[39] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[40] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," *J. Machine Learning Research*, vol. 9, pp. 759–813, Sep. 2008.

[41] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[42] W. Hachem, P. Loubaton, and J. Najim, "Deterministic equivalents for certain functionals of large random matrices," *Ann. Applied Probability*, vol. 17, no. 3, pp. 875–930, Jun. 2007.